DOCUMENT RESUME

ED 064 357 TM 001 525

AUTHOR Kolakowski, Donald; Bock, R. Darrell
TITLE Maximum Likelihood Estimation of Ability Under the Normal Ogive Model: A Test of Validity and an Empirical Example.
SPONS AGENCY National Science Foundation, Washington, D.C.
PUB DATE Apr 72
NOTE 21p.; Paper presented at the annual meeting of the American Educational Research Association (Chicago, Illinois, April 1972)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Ability Identification; *Guessing (Tests); *Mathematical Models; *Measurement Instruments; Models; Psychometrics; *Statistical Studies; *Test Results; Test Validity

ABSTRACT

The estimation of values of a latent-trait presumed to underlie a given set of item response data is made on the basis of dichotomously scored items utilizing the so-called "normal ogive model" of Lawley and Lord. This model provides an internal scale of measurement, scores which are independent of the particular test items employed, individual estimates of the standard error of each subject's score, and a statistical test of how well the data conforms to the constraints of the model. Tables and graphs present the study results. (Author/DB)

ED 064357

U.S. DEPARTMENT OF HEALTH
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

MAXIMUM LIKELIHOOD ESTIMATION OF ABILITY UNDER THE NORMAL OGIVE MODEL: A TEST OF VALIDITY AND AN EMPIRICAL EXAMPLE.

Donald Kolakowski, University of Connecticut

R. Darrell Bock, University of Chicago

Paper presented at the annual meeting of the American Educational Research Association in Chicago, April 3-7, 1972.

TM 001 525

This investigation was supported in part by National Science Foundation Grant GJ-9 to the University of Connecticut Computer Center and NSF Grant GS-1930 to the University of Chicago.

I. Introduction

Physical and physiological measurements are not generally subject to the limitations inherent in psychological testing, where an unknown range of individual variation is compressed into a relatively restricted distribution of scores from a typically 10- to 40- item test. Such psychometric variables produce raw scores distributions which tend to be skewed and platykurtic, their particular properties being dependent upon the difficulty and discriminating power of the test items employed (Lord & Novick, 1968, pp. 386-392). To make valid inferences about the nature of these quantitative traits, especially by means of distributional analyses, it is apparent that we need mental variables possessing better metric properties than is usually the case. A theoretical solution for the hypothetical value of a trait or ability presumed to underlie a given set of item-response data is provided by the latent-trait psychometric models (Lord & Novick, 1968, Chs. 16-20; Rasch, 1960) In the present study we consider the estimation of these trait values on the basis of n dichotomously-scored items utilizing the so-called "normal ogive model" of Lawley and Lord (Lawley, 1943; Lord, 1952; Bock & Lieberman, 1970). This model provides an internal scale of measurement, scores which are independent of the particular test items employed, individual estimates of the standard error of each subject's score, and a statistical test of how well the data conforms to the constraints of the model.

2. The Normal Ogive Model

Consider an unobservable, continuous variable, θ, the "latent ability" of the subjects, which is distributed normally in the population of reference with a mean 0.0 and variance 1.0. Letting $r_{ij}=1$ indicate a correct response by subject i to a dichotomously scored item j, and $r_{ij}=0$ otherwise, define

$$P_{ij} = \text{Prob}\left\{r_{ij} = 1\right\} \qquad (1)$$
$$= \Phi\ (c_j + a_j\ \theta_i)$$

where Φ is the cumulative normal distribution function,

c_j is an index of the difficulty of item j

and a_j is an index of the discriminating power of item j.

Then if $\underline{v}_i = [r_{ij}]$, the n x 1 score vector for a given subject, with ability θ_i,

$$P(\underline{v}_i) = \prod_{J=1}^{n} P_{ij}^{rij}\ Q_{ij}\ (1 - r_{ij}), \text{ where } Q_{ij} = 1 - P_{ij}, \qquad (2)$$

on the assumption of "local independence"; ie that the probabilities of a correct response to any two items for a given value of θ are statistically independent of each other. (They are necessarily independent of θ since θ does not vary.)

A discussion of the plausibility of the normal ogive response characteristic can be found in Lord and Novick (1968, Ch. 16). However, the adequacy of the model must be verified for a given sample of test data. A common situation in which one would not expect a good fit is that in which subjects guess at unknown answers, thus raising the lower asymptotic value of P_{ij} considerably above zero. Equation (1) is easily generalized to include this possibility:

$$P_{ij} = g_j\ [1 - \Phi\ (c_j + a_j\theta_i)]\ + \Phi\ (c_j + a_j\theta_i) \qquad (3)$$
$$= g_j + (1 - g_j)\ \Phi\ (c_j + a_j\theta_i)$$

where g_j is a constant specifying the probability of a chance correct response to item j when the answer is unknown.

In general, the model requires that

(a) the test in question is measuring substantially one trait (ie. a unifactor test)

(b) the probability of answering a given item correctly increases monotonically with the subject's level on the trait, and

(c) the principle of local independence given above.

3. Maximum Likelihood Estimation of Latent Ability and Item Parameters

The estimation of the parameters of the model may be approached from an unconditional or conditional point of view, depending upon whether the subjects are regarded as having been sampled from a specified population or are treated as given entities (see Bock, 1972). The former approach has proven to be extremely time consuming for tests of more than, say, ten items (Bock & Lieberman, 1970). The latter leads to simultaneous estimation of both subject and item parameters and has been adopted here because of its computational efficiency.

A. Estimating ability when the item parameters are known

Letting the parameters of the model be defined as in section 2, and P_{ij} be defined by equation (3), $P(\underline{v}_i)$ in equation (2) is the likelihood function of θ for a given subject. Omitting the i and j subscripts for convenience,

$$\ell = \log P(\underline{v}) = \Sigma r \log P + \Sigma(1-r) \log Q \qquad (4)$$

Letting $Y_j = c_j + a_j\theta$ and $h(Y_j)$ = the unit normal ordinate.

$$\frac{\partial \ell}{\partial \theta} = \Sigma\left(\frac{r}{P} - \frac{1-r}{Q}\right) \frac{\partial P}{\partial \theta} = \Sigma \frac{r-P}{PQ} a\,(1-g)\,h\,(Y) = 0.$$

Also, $\frac{\partial^2 \ell}{\partial \theta^2} = \Sigma \frac{a^2 (1-g) h(Y)}{PQ} \left[(r-P) \left(-Y - \frac{(1-g)h(Y)}{P} + \frac{(1-g)h(Y)}{Q}\right) - (1-g)h(Y) \right]$

since $\frac{\partial^2 \ell}{\partial \theta^2} = a^2 (1-g)h'(Y)$ and $\frac{h'(Y)}{h(Y)} = -Y$.

Applying the Newton-Raphson method to any k-th stage estimate of θ,

$$\theta_{k+1} = \theta_k - \left(\frac{\partial \ell}{\partial \theta}\right) / \left(\frac{\partial^2 \ell}{\partial \theta^2}\right).$$

In the absence of guessing, of course, all computations are performed with the g_j set equal to zero.

B. Estimating item parameters when ability is known.

Given the values of θ, subjectis of similar ability can be grouped to provide an empirical estimate of the proportion of correct responses to each item, at intervals along the ability continuum. Item parameters can then be estimated by means of probit analysis. (Finney, 1971; Bock and Jones, 1968). This solution is presented in detail in Kolakowski & Bock (1970)

C. Estimating ability and item parameters simultaneously.

The above solutions for each set of parameters are developed in terms of the other set. A computer program has been developed to estimate each set in turn, iterating until convergence is reached. (Kolakowski & Bock , 1970), Four to six estimation cycles usually produce stable values. Because the origin and unit of measure are abitrary, the subject parameters are standardized to zero mean and unit variance and all items are calibrated relative to the metric.

The g_j are presently treated as constants which must be determined by inspection. Subjects for whom the procedure will not converge are assigned a default value and, in the present investigation, are eliminated from subsequent analysis. The number of groups or fractiles used in partitioning the subjects for the probit analysis is arbitrary.

4. The Problem of Bias in the M.L. Estimate of Ability

A. Generation of synthetic item responses.

Recall that

$$P_{ij} = \text{Prob}\left\{r_{ij}=1\right\} = g_j + (1 - g_j)\Theta(c_j + a_j\,\theta_i)$$

Assuming constant values for the four parameters of the model, synthetic response data can be generated by sampling a number n_{ij} between 0.0 and 1.0 from the rectangular distribution and assigning the values

$$r_{ij} = 1 \text{ for } n_{ij} \leq p_{ij}$$

$$= 0 \text{ otherwise}$$

This algorithm was performed using 38 previously calibrated test items, for values of $g_j = 0.0$ and 0.15, and a sample of 750 random normal deviates θ_i, hereafter referred to as "true scores." Estimates of these true scores, $\hat{\theta}$; and of the original item parameters, $\hat{a}_j$ and $\hat{c}_j$, were then recovered from both sets of response data using 20 fractiles and an empirical prior. Execution time for runs of six complete estimation cycles on an IBM 360/65 computer was under 4.5 minutes.

B. Comparison of distributional forms without guessing.

For maximum sensitivity to the distributional forms, five tests of normality were employed: the coefficients of skewness and kurtosis, the U-statistic

= ratio of sample range to std. dev. (David et al , 1954), Geary's A = ratio of mean deviation to std. dev. (Geary, 1947), and a Chi Square test on 18 degrees of freedom. Table 1 presents these indices for the distribution of true scores and that of the resultant raw scores, thus illustrating the unacceptable properties of the latter.

Table 2 (a) presents our results for the recovered estimates assuming the $g_j = 0.0$, ie. no chance responses. Although all of the ability distributions have a mean of zero and variance of one by construction, the form of the distribution of the θ_i is leptokurtic and skewed to the right (Fig.1), indicating that subjects of high ability receive inflated trait estimates. This is explained by referring to the graph of original vs. recovered item parameters (Figure 2), in which it is apparent that the easiest and most discriminating items are estimated as being even more extreme, thus defining a lower bound for ability, but having little weight in most calculations. On the other hand there is very little bias in the $\hat{a}_j$ and $\hat{c}_j$ of easy items. The net result is a relative contraction of the left tail of the distribution.

A systematic correction for such asymmetrical bias is difficult to conceive. However, the loss of a small number of unrealistically extreme subjects in the context of a distributional analysis can be tolerated. Therefore, since there were no true scores beyond the range of approximately $\pm$ 3 standard deviations, we accordingly removed the five subjedts whose trait estimates had an absolute value greater than 3.0. Tabel 2(b) shows that the distribution of remaining subjects does not significantly differ from normality on any of the five indices.

Similar analyses were performed for subtests of 10 and 20 items,

selected to uniformly span the entire range of difficulty. The program failed to converge to stable parameter estimates for a 10-item test. Apparently, this is too few items to adequately describe an underlying normal distribution, even with such a large number of subjects, and thus confirms the futility of unconditional estimation with only a handful of items (see Bock, 1972). The results for a 20-item test were similar to those for the 38 items (Table 2(c)), with a stronger upward bias than was the case for the longer test. Hence, the possibility exists that the use of large item pools could itself improve the validity of ability estimates.

Given our priviledged knowledge of the true score distribution, the original analysis was performed again assuming a normal prior rather than the usual empirical prior. It can be seen in Table 2(d) that the results for the two approaches are virtually identical. This is not surprising because, whereas the normal prior fits the data more precisely, the extreme cases (in both tails) are given considerably more weight than the moderate subjects.

Lastly, an analysis was performed assuming, contrary to fact, that the "subjects" might have been guessing. Here the procedure failed to converge for each of three reasonable sets of guessing constants, each subjectively determined from an examination of the item response proportions in the 20 fractiles. This tends to indicate that any results obtained under the guessing model when this assumption is unwarranted will undoubtedly be invalid.

C. Comparison of distributions for data containing chance responses 8

The results of the analysis of the synthetic guessing data are presented in Table 3 and Figure 3 for both the guessing and conventional options of the computer program. Whereas removing the extreme $\hat{\theta}_i$ from these distributions eliminated the original leptokurtosis, they remain significantly skewed to the right although not nearly as extreme. Comparing the two response models in Table 3 reveals that the guessing analysis is decidedly less skewed, and therefore more valid, when chance responses are in fact present in the data. However, the sensitivity of these tests will be better appreciated by referring to Figure 3 for a subjective evaluation of the differences between models.

In conclusion, the normal ogive guessing model should be employed when chance responses are likely to be present in the data, but failure of the conventional model to converge for guessing data —and not vice versa— indicates that the procedure will have the most validity in applications where guessing can be ruled out. In any case, if the present methodology is found to be valid for a variety of prior distributions, provision of suitable default values for unrealistically high ability estimates and the use of subjectively determined guessing constants might still allow generation of pools of calibrated items and the implementation of sequential item testing under consistent, if not ideal conditions.

5. Resolution of a Spatial Visualization trait distribution into normal components.

An empirical problem with data meeting the above ideal criteria involved making an inference about the mode of inheritance of an educationally important mental trait, spatial visualizing ability, by contrasting the properties of the separate ability distributions for the sexes. A 29-item audio-visual version of the Guilford-Zimmerman (1953) Spatial test was administered to a sample of 727 eleventh-grade students. The Normal Ogive latent ability estimates were obtained under the conventional model and the forms of the distributions were analysed for the sexes separately. Table 4 (a & b) shows the results after removing extreme cases. Our first-hand knowledge of the data plus the fact of convergence of the parameter estimates under the conventional model, lead us to place considerable faith in the validity of this analysis.

A maximum likelihood decomposition of these distributions into normal components by the method of Day (1969) yield the results in Table 5 (a & b), namely an upper component comprising 51% of the variation in boys' spatial

ability which corresponds to a similar component comprising only 20% of the variance for girls. Given the range of ability estimates from -2.0 to +3.0, the means of .80 and .68 of these components, respectively, are virtually equal. To deal objectively with the significance of the findings, a likelihood ratio χ^2 test on 2 degrees of freedom was calculated to test the fit of only one component. Also, a Pearsonian χ^2 on 16 degrees of freedom was used to check the adequacy of a bimodal model. These indices (Table 5) verified that the deviation from normality shown in Table 4 is due to the presence of two and only two underlying components. This structure is illustrated in Figure 3 against the background of the frequency histograms for the data.

The existence of a sex-differentiating duality in the distribution of a continuous human variable is compelling evidence for a sex-influenced major gene. The above proportions immediately suggest an X-linked recessive allele with frequency close to 0.5. Moreover, the (assumed) common variance of the components for girls is estimated at nearly one half the magnitude of that for boys (Table 5), suggesting an averaging effect in females which does not occur in males. This is entirely consistent with the hypothesis of sex-linkage and decidedly reinforces the correlational evidence for this model. (Stafford, 1961; Hartlage, 1970; Kolakowski, 1970)

6. Discussion

While the importance of a latent-trait measurement model for validly investigating the mode of inheritance of an intellectual ability is apparent, it is equally clear that we need to be able to objectively select one of several conflicting models without resort to considerations external to the estimation problem. Internal corrections for bias and/or the simultaneous estimation of

guessing parameters for the Normal Ogive model are two as yet unrealized approaches which would correct the weaknesses in the present investigation. On the other hand, other psychometric models based upon the logistic distribution may be more promising in this regard. (see Birnbaum, 1968; Rasch, 1960).

For instances in which chance responses can be eliminated on external grounds, the assumption of normality of the components is still open to scrutiny. Lord (1960) has shown that errors of measurement cannot be assumed to be normally distributed if a subject's score is taken to be the number of items answered correctly. Latent traits being continuous and unbounded, however, this assumption is at least plausible. It therefore remains to investigate the bias of the foregoing procedures for a variety of true score distributions or better yet, to specify theoretically the conditions under which unbiased estimates can be expected to obtain.

References

Birnbaum, A. (1968) Some latent trait models and their use in inferring an examinee's ability. Part V in Lord & Novick (1968) Statistical Theories of Mental Test Scores, below.

Bock, R.D. (1972) Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37; in press.

Bock, R.D. and Jones, L.V. (1968) The Measurement and Prediction of Judgement and Choice. San Francisco: Holden Day.

Bock, R.D. and Lieberman, M. (1970) Fitting a response model for n dichotomously scored items. Psychometrika, 35; 179-197.

David, H.A., Hartley, H.O., and Pearson, E.S. (1954) The distribution of the ratio, in a single normal sample, of range to standard deviation. Biometrika, 41; 482-493.

Day, N.E. (1969) Estimating the components of a mixture of normal distributions. Biometrika, 56; 463-474.

Finney, D.J. (1971) Probit Analysis, 3rd ed. Cambridge University Press.

Geary, R.C. (1947) Testing for normality. Biometrika, 34; 209-242.

Guilford, J.P. and Zimmerman, W.F. (1953) Spatial visualization, form B. Part VI of the Guilford-Zimmerman Aptitude Survey. Beverly Hills: Sheridan Supply Company.

Hartlage, L.C. (1970) Sex-linked inheritance of spatial ability. Perceptual and Motor Skills, 31; 610.

Kolakowski, D. (1970) A behavior-genetic analysis of spatial ability utilizing latent-trait estimation. Unpublished Ph.D. dissertation, Department of Education, University of Chicago.

Kolakowski, D. and Bock, R.D. (1970) A Fortran IV Program for maximum likelihood item analysis and test scoring: Normal Ogive Model. Research Memorandum No. 12, Educational Statistics Laboratory, University of Chicago.

Lawley, D.N. (1943) On problems connected with item selection and test construction. Proceedings of the Royal Society of Edinburgh, 61; 273-287.

Lord, F.M. (1952) A theory of test scores. Psychometric Monograph No.7.

Lord, F.M. (1960) An empirical study of the normality and independence of errors of measurement in test scores. Psychometrika, 25; 91-104.

Lord, F.M. and Novick, M.R. (1968) Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley.

Rasch, G. (1960) Probabilistic Models for some Intelligence and Attainment Tests. Copenhagen: Danish Institute for Educational Research.

Stafford, R.E. (1961) Sex differences in spatial visualization as evidence of sex-linked inheritance. Perceptual and Motor Skills, 13; 428.

TABLE 1

38 vocabulary items — 750 subjects — Random data

	Model	Sample	N	Skewness	Kurtosis	U	Geary's A	$\chi^2(18)$	Description
a)		True scores	750	-.075	2.85	5.97	.802	12.3	
b)		Raw scores	750	-.399**	2.61**	4.97**	.812*	130**	left skew platykurtic

TABLE 2

a)	NRMØJ rectangular prior	Ability estimate	749	.384**	4.20**	8.01**	.772**	27.2	right skew leptokurtic
b)	NRMØJ rectangular prior	Ability estimate: extremes removed	744	.125	3.05	5.84	.791	14.6	
c)	NRMØJ rectangular prior	20 items	745	.787**	4.79**	6.98	.756**	43.4**	right skew leptokurtic
d)	NRMØJ normal prior	Ability estimate	749	.338**	4.18**	8.03**	.772**	21.7	right skew leptokurtic

* desigantes $p < .05$

** designates $p < .01$

TABLE 3

	Model	Sample	N	Skewness	Kurtosis	U	Geary's A	$\chi^2(18)$	Description
a)	NRMØJ	Guessing data	750	.773**	4.02**	7.46*	.786	57.9**	right skew leptokurtic
b)	GUESS	Guessing data	740	.347**	3.62**	8.37**	.780**	32.7*	right skew leptokurtic
c)	NRMØJ extremes removed	Guessing data	744	.507**	2.86	5.32*	.802	49.8**	right skew
d)	GUESS extremes removed	Guessing data	737	.283**	2.94	5.63*(?)	.791	32.0*	right skew

* designates $p < .05$

** designates $p < .01$

TABLE 4

	Model	Sample	N	Skewness	Kurtosis	U	Geary's A	$\chi^2(18)$	Description
a)	NRMØJ	Boys	366	.052	2.62*	5.11*	.811	10.7	platykurtic
b)	NRMØJ	Girls	345	.644**	3.36	5.87	.804	42.2**	right skew

TABLE 5

Decomposition of Spatial Data

	Model	Sample	N	Proportions of mixture 1 vs. 2	1	2	2	L.R. $^2(2)$	Pearsonian $^2(16)$
a)	NRMØJ	Boys	366	49% : 51%	-.37	.80	.60	248**	11.7 NS
b)	NRMØJ	Girls	345	80% : 20%	-.60	.68	.32	235**	21.2 NS

* designates $p < .05$

** designates $p < .01$

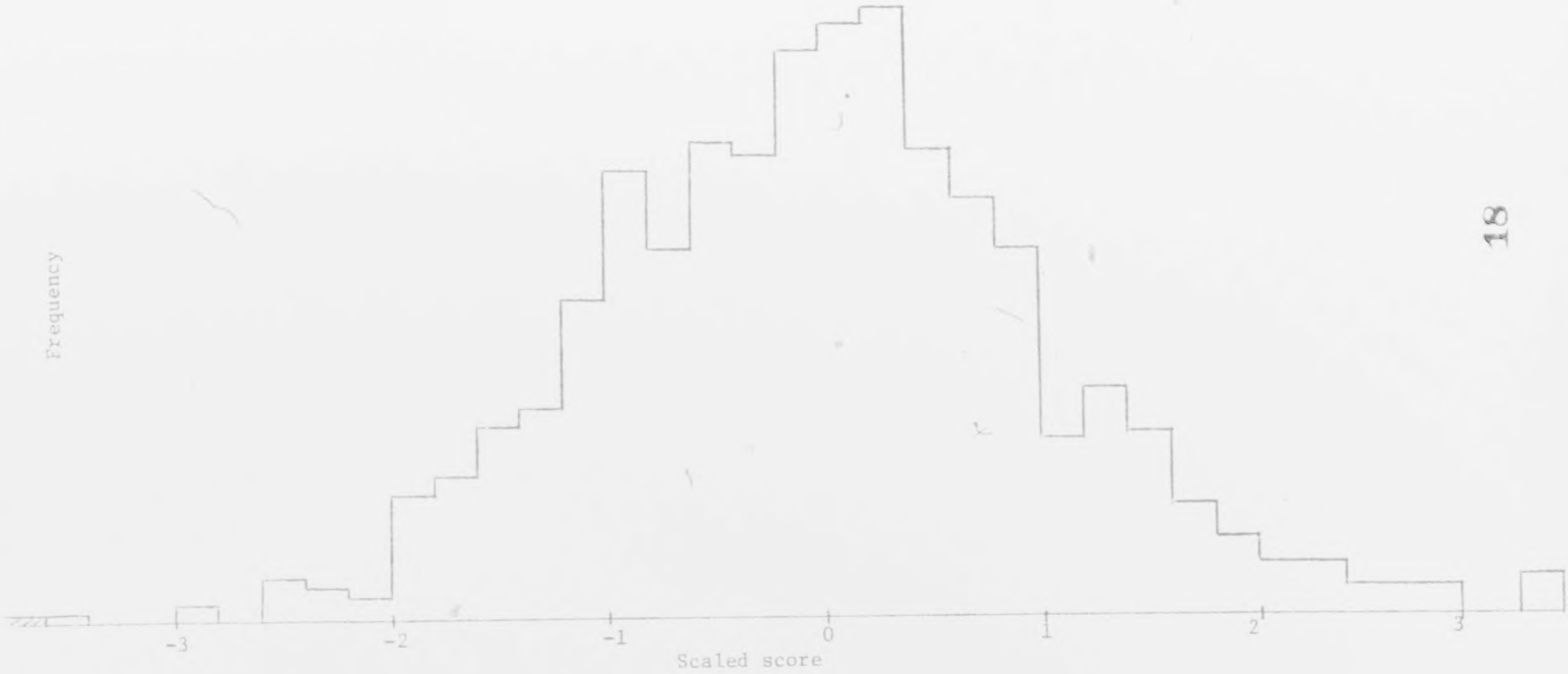

FIGURE 1: Distribution of recovered estimates for the conventional model

FILMED FROM BEST AVAILABLE COPY

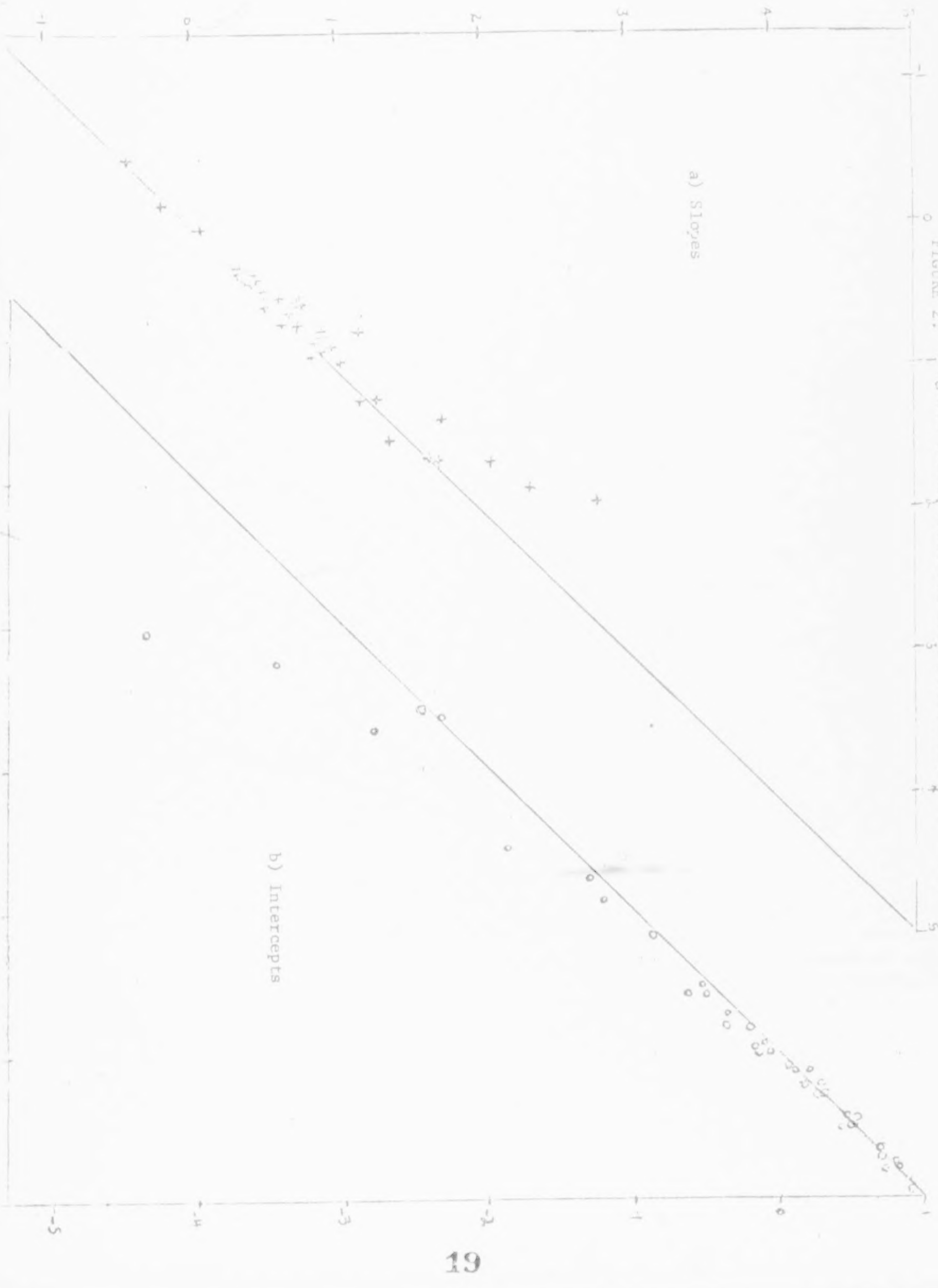

FIGURE 2: Original versus recovered parameters, convolutional model

FIGURE 3: Distributions of recovered estimates from synthetic seismic data

FIGURE 4 : SPATIAL VISUALIZATION ABILITY SCORE DISTRIBUTIONS